不确定 PAHT 聚类算法在滑坡危险性预测上的应用 *

胡　健[a]，朱　玲[b]，毛伊敏[b]

(江西理工大学 a. 应用科学学院 信息工程系；b. 信息工程学院，江西 赣州 341000)

摘　要：针对滑坡预测聚类研究中由于难以确定传统聚类算法需要预先设置的簇个数和无法精准衡量不确定因素降雨量导致预测效果欠佳的问题，提出一种新的聚类算法—不确定 PAHT（partition algorithm on the hierarchical thinking）算法，该算法引入一种不确定数据模型——M-D 距离，其有效刻画了不确定的雨量数据；并结合层次聚类思想，通过找出最佳阈值 p*自动确定 k 值。以延安宝塔区为实例进行对比实验，实验结果验证了不确定 M-D 距离和 PAHT 算法的有效性及不确定 PAHT 算法在滑坡危险性预测上的可行性。

关键词：不确定数据；聚类算法；危险性预测；滑坡

中图分类号：TP399　　doi: 10.3969/j.issn.1001-3695.2017.11.0744

Uncertain PAHT clustering algorithm in landslide hazard prediction application

Hu Jian[a], Zhu Ling[b], Mao Yimin[b]

(1. *Dept. of Information Engineering*, *College of Applied Science Jiangxi University of Science & Technology*, *b. School of Information Engineering Jiangxi University of Science & Technology*, *Ganzhou Jiangxi* 341000, *China*)

Abstract: In the clustering study of landslide prediction, the difficulties of determining the number of clusters which traditional clustering algorithm needs to set in advance and accurately measuring the important factor of Landslide induced-rainfall leads to bad prediction effect. Therefore, this paper proposes a new clustering algorithm-Uncertain PAHT algorithm , the algorithm introduces a kind of uncertain data model called M-D distance, which effectively measure the uncertain rainfall; and based on the hierarchical clustering thinking, through finding the best threshold p* to determine the k value. Contrast experiment in Yenan Baota district as an example, the experimental results verified the effectiveness of uncertain M-D distance and PAHT algorithm and the feasibility of uncertain PAHT algorithm on the landslide hazard prediction.

Key words: uncertain data; clustering algorithm; hazard prediction; the landslide

0　引言

滑坡是一种常见的地质灾害。近年以来，滑坡事件发生的频率和强度均成增长之势，所造成的人员伤亡和经济损失也不断加大，所以如何预防滑坡灾害已成为亟待解决的问题。滑坡预测是有效预防滑坡灾害的重要途径，但滑坡的主要诱发因素—降雨量的不确定性，给滑坡预测研究增加了一定难度，因此不确定数据的分析研究成为重点。

聚类分析是数据挖掘研究中一种常用的分析方法，其主要功能是将数据集中相似的对象尽可能划分在相同的簇，而把相异的对象尽可能划分到不同的簇。聚类算法已广泛用于多个领域，在滑坡研究领域，聚类分析也已多次被研究使用，并取得了一定成果。张俊等人[1]选取 7 个致灾因子作为滑坡易发性的评价指标，采用 K-means 聚类算法对三峡库万州区滑坡易发性评价体系进行分级，实验结果表明滑坡灾害易发性评价体系预测精度较高。郭靖等[2]首先在黔西玄武岩地区建立 Logistic 模型寻找主要致灾因子，其次利用聚类算法对八个致灾因子进行预测，编制滑坡易发性区划图，结果验证了区划结果的可靠性。胡畅等人 [3]以秭归到巴东段的顺层滑坡为研究对象，采用两步聚类算法将库水、降雨等影响因子进行等级划分，然后根据监测数据进行观测分析，取得了较好的预测结果。夏元友[4]提出并建立了一种系统加权聚类算法，在类间距计算时考虑了各因素的影响权重，并以三峡库岸研究程度较高的边坡为例，进行了一般系统聚类法与系统加权聚类算法的对比实验，对比结果表明，其预测精度有较大提高。虽然传统聚类算法在滑坡预测上取得了一定成效，但依然存在以下较明显的问题：首先，对于滑坡的主要诱发因素—降雨量，由于它是不确定数据，传统聚类算法无法对它有效分析和处理。其次，传统聚类算法往往

收稿日期：2017-11-19；**修回日期**：2017-12-28　**基金项目**：江西省教育厅科技项目（GJJ151528，GJJ151531）；国家自然科学基金资助项目（41562019，41530640）；江西省自然基金资助项目（20161BAB203093）

作者简介：胡健（1967-），男，江西赣州人，教授，博士，主要研究方向为数据挖掘、软件工程等（1050023437@qq.com）；朱玲（1994-），硕士研究生，主要研究方向为数据挖掘；毛伊敏（1970-），女，教授，博士，主要研究方向为数据挖掘、地理信息系统等.

需要预先给定簇个数，并以此为终止条件进行聚类，但在滑坡预测应用实例中，无法预先给定 k 值。

针对这些问题，本文提出了相应的改进方法。首先，本文提出一种新的不确定数据距离公式——M-D 距离，该公式是在 Hausdorff 距离的基础上引申出的适用于所有区间数的距离公式，它能更精确的描述两个不确定数据之间的距离。其次，本文提出一种可以有效处理不确定数据的新聚类算法—PAHT 算法，以 COPS（clusters optimization on preprocessing stage）[5]思想为基础，该算法首先引入不确定数据模型—M-D 距离；其次借助层次聚类思想自适应的找出参数 p*；然后以最佳聚类质量对应的阈值 p*为条件，再做一次划分聚类；最后剔除噪声和离群点，得到最佳聚类个数 k 和最终聚类。最后，本文以延安宝塔区为实例进行实验，通过在相同不确定数据处理方式下 PAHT 算法和其他几种典型算法的对比，验证 PAHT 算法的有效性，通过在同一 PAHT 算法下不同不确定数据处理方式的实验结果对比，验证 M-D 距离的有效性，两组实验均验证了不确定 PAHT 算法在滑坡预测研究上的可行性。

1　不确定数据处理

本章首先介绍不确定数据定义，提出区间数的概念，其次引出一种新的不确定数据间的距离公式 M-D 距离，最后通过以 M-D 距离构建一个新的排序函数提出一种新的不确定数据排序方法。

1.1　不确定数据定义

不确定数据的表示方式有多种，例如：决策数据的三角模糊数，传输数据的点概率数、测量数据的区间数等[6]。本文研究的不确定数据为区间数，区间数用区间的形式来表示数据的不确定性，其定义如下[7]:

定义 1　给定 $A^-,A^+\in R^d$，且 $A^+\geq A^-$，称集合 $A=[A^-,A^+]$ 为一个区间数，其中 A^- 为区间数 A 的下限，A^+ 为区间数 A 的上限。当 $A^-=A^+$ 时，即上下限相等时，区间数 A 为一个精确数。

1.2　不确定数据距离

在聚类算法中，距离是个非常重要的概念，聚类算法通常采用距离作为相似性的评价指标，即认为两个数据之间距离越近，其相似度就越大。在确定性数据中，欧式距离是应用最广泛的度量空间，它能最直观反映两个点之间的真实距离，但在不确定性数据中，欧式距离无法有效度量其间的距离，为描述其不确定性，文献[8]提出了基于区间数的 Hausdorff 距离：

定义 2　设两个区间数 $A=[A^-,A^+]$ 和 $B=[B^-,B^+]$，式中 $c(x)$ 表示区间数 x 的中点；$r(x)$ 表示区间数 x 的半径（$x=A$，B），则两点之间的 Hausdorff 距离为：

$$H(A,B)=|c(A)-c(B)|+|r(A)-r(B)| \tag{1}$$

分析式（1）易看出，在均匀分布的前提下，式中 $c(x)$ 能直观刻画区间数的集中位置，而半径 $r(x)$ 则能有效反映区间数的离散程度。但是通常区间内点数据的分布情况往往无法获知，基于此，本文借鉴 Hausdorff 距离思想提出一种新的适用于所有区间数的 M-D 距离：

定义 3　设两个区间数 $A=[A^-,A^+]$ 和 $B=[B^-,B^+]$，式中 $M(x)$ 表示区间数 x 的均值；$D(x)$ 表示区间数 x 的平均差（$x=A$，B），则两点之间的 M-D 距离为

$$M-D(A,B)=|M(A)-M(B)|+|D(A)-D(B)| \tag{2}$$

注：假设一个区间数 ，区间数的内部点不存在任何分布，其中：

$$M(X)=\frac{X_1+X_2+X_3+\ldots+X_n}{n},$$

$$D(X)=\frac{\sum|X_i-M(X)|}{n}$$

分析式子可得，对于任意区间数，均值 $M(x)$ 能有效反映数据点的集中趋势，而以均值作为参考系的平均差 $D(x)$ 较精准的反映数据的离散程度。

针对 Hausdorff 距离仅适用于均匀分布区间数的局限性，新提出的 M-D 距离广泛适用于一般区间数。在实际应用中，区间数的内部点可能服从任何分布或者不服从分布，相较于传统 Hausdorff 距离会造成部分数据信息丢失导致无法准确度量相似性的情况，基于均值和平均差的 M-D 距离可以充分利用区间内的有效信息，有效度量其相似性。

下面给出 M-D 距离是一个度量空间的证明：

证明　区间数集用 γ 表示，设三个区间数 $A,B,C\in\gamma$，分别证明其间距离满足非负性；对称性；三角不等式性；

1）非负性

$M-D(A,B)=|M(A)-M(B)|+|D(A)-D(B)|$，其中 $|M(A)-M(B)|\geq 0$ $|D(A)-D(B)|\geq 0$，所以 $M-D(A,B)\geq 0$，满足非负性。

2）对称性

$|M(A)-M(B)|=|M(B)-M(A)|$，同样 $|D(A)-D(B)|=|D(B)-D(A)|$，故 $M-D(A,B)=M-D(B,A)$，满足对称性。

3）三角不等式性

$|M(A)-M(B)|+|M(B)-M(C)|\geq$ $|M(A)-M(B)+M(B)-M(C)=|M(A)-M(C)|$；同样地 $|Q(A)-Q(B)|+|Q(B)-Q(C)|\geq$ $|Q(A)-Q(B)+Q(B)-Q(C)=|Q(A)-Q(C)|$ 所以 $M-D(A,B)+M-D(B,C)\geq M-D(A,C)$，满足三角不等式性。

因此 M-D 距离满足度量空间定义的三个条件，是距离度量公式。

1.3　不确定数据排序

由于本文提出的不确定 PAHT 聚类算法为提高算法效率，首先会在属性值上做一次排序以便之后的顺序扫描，所以本节在提出的 M-D 距离的基础上，构建一种新的排序函数，以相应排序函数的值来反映区间数的大小进行排序。下面给出区间数

排序规则[9]

定义 4　设 n 个区间数 $x_1,x_2,x_3,\ldots x_n$， $x_i=[x_i^-,x_i^+]$，找到其最小目标数 Min，$Min=\inf\left(\bigcup_{i=1}^{n}S(x_i)\right)$，其中 $S(x_i)$ 是区间数 x_i 的支集，在定义 3 的基础上提出排序函数：

$$f(x_i)=|M(x_i)-M(z)|+|D(x_i)-D(z)| \tag{3}$$

其中 z 为最小目标数 Min，由于排序函数以最小目标数为参照系，所以若 $f(x_j)>f(x_k)$ 则可以得出区间数 $x_j>x_k$.

2　不确定 PAHT 算法

本章介绍不确定 PAHT 算法的思想和基本步骤。由于传统聚类算法需要预先设置 k 值且在处理不确定数据上因无法准确衡量其不确定性致聚类效果欠佳，故提出不确定 PAHT 算法，其基本思想是：引入新提出的不确定数据模型—M-D 距离，其次通过在排好序的数据集上做顺序扫描，根据 p 值的逐步增加做不同划分并保存 CF 统计值，增量的构建一条聚类质量曲线，自适应的找出最佳聚类指标值对应的阈值 p*，再以 p*为条件做划分聚类自动确定 k 值，得到最佳聚类效果。

不确定 PAHT 算法设计如下：

a)数据进行预处理，获得有效数据集；

b)将数据集以式（3）进行从小到大排序，形成有序序列；

c)以 M-D 距离为度量，以 $|x_j-y|<p$ 为扫描范围在有序序列上顺序扫描，根据参数 δ 逐步增加 P 值，分别做不同划分，直至所有点聚类到同一个簇中，保存每次划分的 CF 值；

d)根据每次划分保存的 CF 值，增量的绘制聚类质量曲线；

e)取聚类指标极小值点对应的阈值 p^*，并以此为条件进行一次划分，得到 m 个子集；

f)计算 m 个子集每个子集中点的数目，删除点数目少的子集;

g)得到 k 个簇，输出数据集聚类结果。

2.1　快速排序

利用新提出的排序函数把每个区间数化成一个可以代表区间数的实数，其次利用快速排序方法对区间数进行从小到大排序，其基本思想是[10]：通过一趟排序将要排序的数据分割成独立的两部分，其中左半部分的所有数据都比右半部分的所有数据都要小，然后再按此方法对这两部分数据分别进行快速排序，整个排序过程递归进行，直至每个部分只有一个数据，以此达到整个数据变成有序序列。

2.2　相似点和参数设置

定义 5　设阈值 $p_j\geq 0$，$1\leq j\leq d$.若 $|x_j-y_j|\leq p_j$，则称 X 点和 Y 点在第 j 维相似,给出阈值向量 $P=\{p_1,p_2,\ldots,p_d\}$，这里阈值 P 在每个维度的值不等，它反映不同维度属性值的分布情况。若 $|x_j-y_j|\leq p_j$ $(j=1,2,\ldots,d)$，则称 X 点和 Y 点为相似点。

彼此相似的点构成同一个簇。由公式易看出，若阈值向量 P 无限小，每个单独的点构成一个簇，若阈值向量 P 无限大，则所有的点构成一个簇.

算法计算过程从 $P=P^0$ $(P^0=\{0,0,\cdots,0\})$ 开始，逐步增加 P 值，每步增量为 $\delta(\delta=\{\delta_1,\delta_2,\cdots,\delta_d\})$，每一步随着 P 值增加形成不同划分，当原本属于不同簇的点变为相似点，这些簇合并为同一个簇，每步计算其有效性指标 Q 值（其计算过程将在 2.3 节详细介绍），直至所有点都在一个簇中。而增量 δ 的选取与数据集的分布稀疏度有关，给出相关定义：

定义 6　数据集在第 j 维的分布稀疏度 ω_j 为

$$\omega_j=\sqrt{\frac{\sum_{i=1}^{n}\left(x_{ij}^*-\theta_j\right)^2}{n-1}}$$

其中: x_{ij}^* 为数据点 X_i 在第 j 维的 [0,1] 规范化值，而 θ_j 则表示第 j 维的中心，即

$$x_{ij}^*=\frac{x_{ij}-\min_{l=1,\ldots,n}\{x_{lj}\}}{\max_{l=1,\ldots,n}\{x_{lj}\}-\min_{l=1,\ldots,n}\{x_{lj}\}}\qquad ,\theta_j=\frac{1}{n}\sum_{i=1}^{n}x_{ij}^*$$

可以看出 ω_j 就是数据集第 j 维规范化的标准偏差，其值反映第 j 维属性值的分布稀疏程度，可以通过维度上属性的不同来查找可能存在的簇结构，故可以推断出参数 δ 公式：

$$\delta_j=\varepsilon\times\frac{\max\{\omega_1,\omega_2,\ldots,\omega_d\}}{\omega_j}$$

其中：ε 取值经过实验反复验证，设定为 $\varepsilon=0.01$.

2.3　聚类质量曲线

算法旨在寻找聚类质量曲线中最佳聚类质量对应的阈值 P*，聚类质量通常利用聚类有效性指标来衡量，一个质量好的聚类结果的基本特征是相同簇内的数据点尽可能相似，不同簇内的数据点尽可能相异，有效性指标量化这种相似度和相异度并组合二者，本算法选用有效性指标 Q[5]。其定义如下：

定义 7　给定一个数据集的划分 $C^k=\{C_1,C_2,\ldots C_k\}$，$Scat(C^k)$ 表示 C^k 的簇内相似度，$Sep(C^k)$ 衡量 C^k 的簇间相异度，$Q(C)$ 表示为

$$Q(C)=\frac{1}{M}\left(Scat(C)+Sep(C)\right)$$

当 $k=1$ 时，$Seq(C^1)=0$、$Scat(C^1)=M$，易求得 $Q(C^n)=Q(C^1)=1$，故最佳聚类指标值 $Q(C^*)$ 取 (0,1) 上的极小值点。描绘聚类质量曲线需计算每次划分的聚类指标值 $Q(C^k)$，通过 $Scat(C^k)$ 和 $Sep(C^k)$ 的增量计算可以简化 $Q(C^k)$ 的计算

$$Sep(C^{k-1})-Sep(C^k)=-2\sum_{j=1}^{d}\frac{LS_{mj}LS_{nj}}{|C_m||C_n|}-$$

$$\sum_{j=1}^{d}\left((k-2)\frac{|C_n|^2SS_{mj}+|C_m|^2SS_{nj}}{(|C_n|+|C_m|)|C_m||C_n|}+\sum_{i=1}^{k}\frac{SS_{ij}}{|C_i|}\right)-$$

$$2\sum_{j=1}^{d}\left(\frac{|C_n|^2LS_{mj}+|C_m|^2LS_{nj}}{(|C_n|+|C_m|)|C_m||C_n|}\sum_{i=1,i\neq m,n}^{k}\frac{LS_{ij}}{|C_i|}\right)<0$$

$$Scat(C^{k-1})-Scat(C^k)=$$

$$2\sum_{j=1}^{d}\left(|C_m|SS_{nj}+|C_n|SS_{mj}+2LS_{mj}LS_{nj}\right)>0$$

基于此，数据集的合并操作可以简化成 $|C_i|$，SS_{il}，LS_{il} 相邻数值间简单的加法计算，故只需为每次划分 C^k 保存一个 CF 统计值[11]：

$$CF_i=(|C_i|,\langle SS_{i1},LS_{i1}\rangle,\langle SS_{i2},LS_{i2}\rangle,...,\langle SS_{id},LS_{id}\rangle)$$

注：在计算初始值 M 时，需获得每个点的 CF 结构

根据此方法可以较快画出聚类质量曲线，下图给出 $Q(C)$ 曲线的示例图 1，最佳聚类指标 $Q(C^k)$ 对应的阈值为 p^*

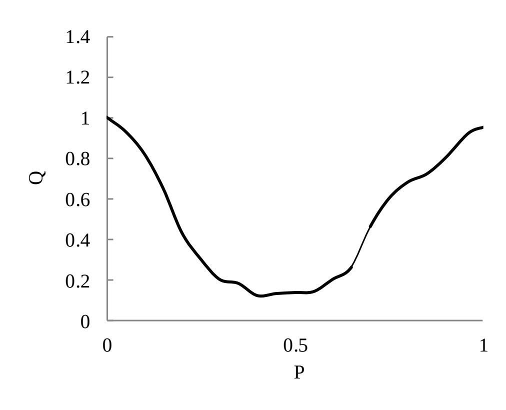


图 1　Q(C)示例

2.4　划分聚类

以上节得到的阈值 p^* 为条件做一次划分聚类自动确定簇的个数，其具体步骤如下：在有序属性列上，以第一个点为 x_1 开始做顺序扫描，其扫描范围为 $|x_1-y|\le p^*$，直至没有点满足这个条件，然后取最后一个满足 $|x_1-y|\le p^*$ 的 y 点为 x_1，仍以 $|x_1-y|\le p^*$ 为扫描范围扫描其后的点，以此循环，直至没有点满足此条件。取之后的点为 x_2，以 $|x_2-y|\le p^*$ 为范围做顺序扫描，重复以上步骤，直至最后一个点也被扫描，所有满足 $|x_k-y|\le p^*$ 的点构成第 k 个簇，所有相似点划分到相同簇中，得到最终聚类结果。

2.5　去除噪声点

在上节划分聚类所得的结果簇中仍存在噪声和离群点，它的基本特点是：簇中可能只包含一个数据点或数据点数目较少。为去除噪声点，在划分聚类的结果上，分别计算其 k 个类中数据点的数目，删除所有数据点数目较少的类，以此得到的最终聚类结果为最佳聚类，得到的簇个数为最佳 k^* 值。

2.6　时间复杂度分析

算法的时间复杂度总共由几部分构成，其中快速排序部分的时间复杂度为 $O(dn\log n)$，d 代表数据的维数;生成不同 Q 值的时间复杂度为 $O(pdnN^*)$,其中 p 为循环次数，N^*代表 δ 邻域内相似点平均数目，它们的值远远小于 n；划分聚类的时间复杂度为 $O(dnN^*)$；去除噪声点的时间复杂度是 $O(n)$,综上所述，PAHT 算法的时间复杂度为 $O(dn\log n)$.

3　实验研究及结果分析

本章以延安宝塔区为实例，提取实验相关数据，分别引进第 2、3 节提出的不确定数据模型和不确定 PAHT 算法，首先以在新提出的 M-D 距离处理不确定数据的基础上，比较 PAHT 算法与其他算法的预测精度，验证 PAHT 算法的有效性，其次以 PAHT 算法为基础，分别使用传统不确定数据处理方式和 M-D 距离处理方式，观察实验结果，证明 M-D 距离在衡量不确定数据上的有效性及不确定 PAHT 算法在滑坡预测上的可行性。实验在 intel-i7 双核、内存 8G 的计算机上运行，操作系统为 Windows 旗舰版。

3.1　数据来源及处理

延安宝塔区的所有数据均来自于西安地质调查中心，首先利用软件 ARCGIS 将宝塔区进行栅格化，选取 $5m\times5m$ 的栅格分辨率，最后得到 5672922 个栅格单元。实验选取七个属性作为滑坡危险性预测的评价因子[12],分别为坡型、坡向、坡高、坡度、岩土体、植被和降雨量。其中，坡型、坡向、坡高和坡度的数据从 1:5000 精度的数字高程图里各因子的专题图中分别获取[13]，岩土体数据从 1：1000 的地质图中获得，植被数据通过 EVNI 遥感软件取得，降雨量值采用滑坡发生前后 7d 的日降雨量区间值。

在这些因子中，坡高、坡度、坡向均为连续属性，可直接进行归一化处理，坡型（凹型，凸型，阶梯型，直线型）、植被（低，较低，高，较高）及岩土体结构（黄土+近于水平古土壤层型，黄土+倾斜古土壤层型，黄土+古土壤+基岩型，黄土+古土壤+新近纪泥岩型）为离散属性，需要先将其数值化再进行归一化处理，而降雨量的表现形式为一个区间数，它具有不确定性，传统方法无法对其有效刻画，故利用本文第 2 章提出的不确定 M-D 距离对其进行处理。

3.2　不确定 PAHT 算法滑坡预测模型的构建

首先引入不确定数据模型，以 M-D 距离衡量不确定数据间的距离，用以 M-D 距离为基础提出的排序函数进行不确定数据间的快速排序，其次基于层次聚类思想增量的构建一条聚类质量曲线，利用参数 δ 逐步增加 p 值做不同划分，每次划分保存 CF 统计值，以此为基础增量的计算聚类指标值 $Q(C^k)$，画出 $p-Q$ 曲线，找出聚类指标 Q 的极小值点对应的 p^* 值，再以最佳阈值 p^* 为条件做一次划分，以 $|x_k-y|\le p^*$ 为范围顺序扫描有序序列，得到划分聚类结果，最后去除噪声和离群点，以此得到的聚类结果为最佳聚类，得到的类别个数为最终簇个数。

3.3　滑坡危险性等级划分

滑坡危险性等级是滑坡危险性预测的决策因子，滑坡危险性等级分为：低危、中危、高危。这里根据上述建立的不确定 PAHT 算法的预测模型，把 5672922 个评价单元最终聚类到 465 个子集中。每个子集中的点具有相似特征，所以可以利用“与发育滑坡的相似特征也同时具有相似的滑坡发生趋势”[14]这一特性，根据已有的 293 个滑坡观测点的已知危险性等级，预测每个聚类子集中点的危险性等级。实验选用直接搜索法和专家评价法[15]，首先使用直接搜索法对每个子集进行扫描，若子集中只有一个确定的危险性等级，则该危险性等级就是这整个子

集的危险性等级，若子集中含有的确定性危险性等级不等，则遵从少数服从多数原则确定子集的危险性等级，对于未含确定危险性等级或含有相同数目的不同危险性等级的聚类子集，它们的危险性等级由专家结合区域调查结果再根据经验进行评定。

3.4 评价标准

实验选取预测精度和 Kappa 系数两个指标作为滑坡预测的评价标准，预测精度即预测点和观测点一致的数量与总体观测点数量的比例，而 kappa 系数作为一种被广泛使用的一致性评价机制，在滑坡预测评价实例中，通过考虑混淆矩阵的所有因子[16]来反映预测结果和观测数据之间的吻合程度，其取值范围为[-1.1].其值越大，表明预测值和观测值的一致性越大.其公式[17]为

$$Kappa = \frac{p_o - p_e}{1 - p^e}$$

$$p_o = \frac{\sum_{i=1}^{n} p_{ii}}{N}$$

$$p_e = \frac{\sum_{i=1}^{n} (p_{i+} \times p_{+i})}{N^2}$$

其中：p_o 是预测精度，表示预测和观测一致的概率，p_e 表示由偶然机会造成的预测点与观测点一致的概率，n 为所有类别数，N 为样本总数，P_{ii} 为第 i 类型被正确分类的数目，P_{i+} 为第 i 类型所在列的数目之和，P_{+i} 为第 i 类型所在行的数目之和。

3.5 滑坡危险性预测结果评价分析

3.5.1 实验 1

为验证 PAHT 算法的有效性，实验抽样不同比例的实验数据，选用其他几种典型算法与 PAHT 算法实验，比较它们的预测精度和时间性能。

取 CFSFDP、SYNC、FAKCS 算法进行对比实验。其中，CFSFDP 算法以密度峰值作为聚类中心，凭借相邻距离和密度完成簇的划分，SYNC 算法则是将数据集中的每个数据点的每个属性看做一个相位振子，随同步范围慢慢扩大，所有振子会慢慢形成多个做局部同步运动的簇，FAKCS 算法基于 Davies-Bouldin 指标自适应查找参数 ε，并能自动确定最佳聚类数。

为验证算法的聚类精度，实验数据分别取所有实例数据中的 0.1%、0.5%、1%、5%、8%和 10%，所有算法都以传统方式处理连续和离散属性的数据，以 M-D 距离公式处理不确定数据，实验结果如图 2 所示。

从图中易看出，PAHT 算法的平均性能较优于其他三种算法。其中 CFSFDP 算法总体聚类质量较差，平均预测在 75%左右，这是因为数据集中存在部分不明显类簇，而 CFSFDP 算法往往不擅长发现此类型簇致聚类效果欠佳，从 SYNC 算法折线可以看出，随数据集变大，其预测精度越低，可以得出 SYNC 算法在大规模数据集的聚类上存在局限性， FAKCS 算法聚类质量虽略优于 CFSFDP 算法和 SYNC 算法，它的平均预测精度为 85%左右，仍低于 PAHT 算法高达 90%的预测精度，根据不同算法的对比，易得出 PAHT 算法的预测效果最佳。

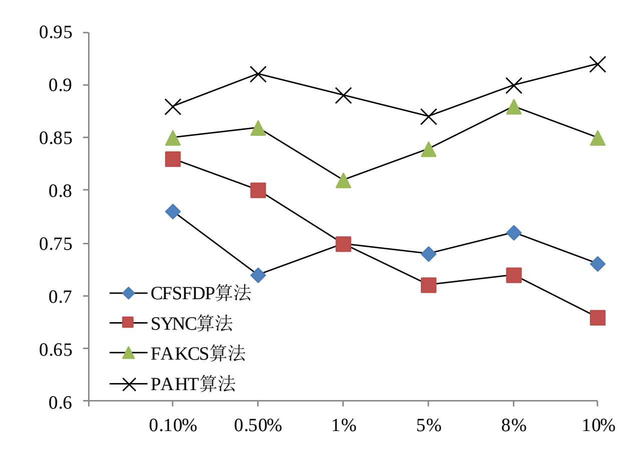


图 2 不同算法预测精度对比结果

为验证算法的时间效率，分别取 0.1%、1%、10%三种抽样比例下各算法的运行时间，实验对比效果如图 3 所示。

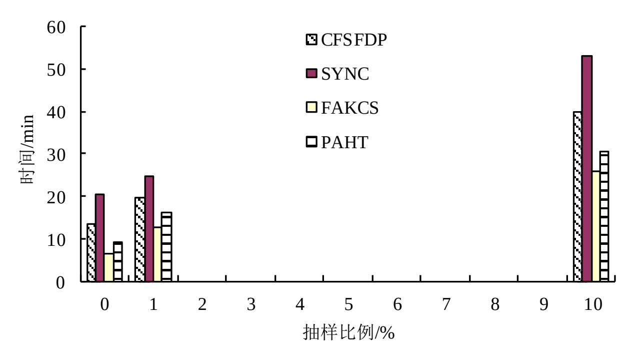


图 3 不同算法运行时间对比结果

从图中可以看出，PAHT 算法和 FAKCS 算法在三种不同抽样比例的数据下时间性能较优。这主要是由于 SYNC 算法需要对每一个分量进行迭代计算，迭代 T 次的时间复杂度为 $O(T \times n^2)$ ， CFSFDP 算法的运行时间虽略低于 SYNC 算法，但它需要计算每个簇的边界区域，时间复杂度达到 $O(n^2)$ ，FAKCS 算法通过对原始数据集进行压缩使算法计算量大大减少，适合处理大数据集，但压缩后的数据集不能充分代表原数据集，所以其预测精度不如 PAHT 算法，而 PAHT 算法通过排序方法提高时间效率，其时间复杂度为 $O(dn\log n)$，具有较好的时间性能。综合预测精度及时间性能的考虑，PAHT 算法的有效性优于其他三种算法。

3.5.2 实验 2

为验证不确定距离 M-D 距离的有效性，实验使用 PAHT 算法，分别以不同方式处理不确定数据，比较它们的危险性等级划分及预测精度和 Kappa 系数。

取 Euclidean 距离和 Hausdorff 距离与 M-D 距离进行对比实验，在滑坡危险性预测的传统聚类算法中，通常以定量方式[18]处理降雨量这一不确定数据，通过雨量值大小将其划分为：小雨，中雨，大雨，暴雨，大暴雨，特大暴雨，以传统 Euclidean

距离衡量两个对象间的距离；Hausdorff 距离广泛用于不确定数据距离衡量，它利用不确定数据的中点和半径来表现其不确定性。实验取宝塔区实例数据，其区域滑坡观测点共 428 个，在数据预处理阶段，所有观测点被栅格化为 1367 个单元，其中含降雨信息单元为 1036 个，其余 331 个点为不含降雨信息的稳定单元，取 PAHT 算法，分别以 Euclidean 距离、Hausdorff 距离和 M-D 距离衡量不确定数据间距，实验结果如表 1 所示。

表 1　不同不确定数据处理方式下 PAHT 算法的滑坡危险性等级划分表

不确定数据处理方式	观测	预测			P_o	Kappa
		低危	中危	高危		
Euclidean 距离	低危	330	72	31	77.90%	0.6490
	中危	37	540	61		
	高危	32	69	195		
Hausdorff 距离	低危	358	50	25	84.19%	0.7505
	中危	24	561	53		
	高危	15	49	232		
M-D 距离	低危	386	36	11	90.24%	0.8464
	中危	18	590	30		
	高危	8	18	270		

由表 1 中可以看出，实验在其他条件均相同的情况下，仅通过不同方式的不确定数据处理，得到的实验结果相差较大，传统定量法通过 Euclidean 距离处理不确定数据，由于它无法刻画其不确定性导致其预测精度不到 80%，Hausdorff 距离虽然综合考虑了不确定数据的不确定性，但它忽略了数据内部点的分布信息，它的预测结果虽高于传统定量方法，但仍未达到预期需求，M-D 距离通过考虑不确定数据的均值和均值差来刻画其不确定性，充分考虑了其内部分布情况，其预测精度达到 90.24%，Kappa 系数达到 0.8464，实验结果证明了它在衡量不确定数据上的准确性。

3.5.3 实验小结

实验 1 通过在相同条件下几种典型算法与 PAHT 算法的对比实验证明了 PAHT 算法的有效性，其综合了预测精度和时间效率两个不同角度。实验 2 取不同不确定数据处理方式在 PAHT 算法上实验，通过对比滑坡危险性等级划分表验证了 M-D 距离在衡量不确定数据中的有效性。同时，两个实验的实验结果都验证了不确定 PAHT 算法在滑坡危险性预测上的可行性。

4　结束语

针对传统聚类算法在滑坡预测应用上对降雨量值刻画困难及无法预先给出 k 值等问题，本文提出一种新的不确定数据距离-M-D 距离，并以此为模型提出不确定 PAHT 算法，该算法通过扫描有序序列增量的构建聚类质量曲线，并自适应的找出最佳阈值 p*，以此为划分自动确定最终聚类 k 的数目。延安宝塔区的实验结果表明，不确定 PAHT 算法在提高滑坡危险性预测精度上取得了较好效果。

参考文献：

[1] 张俊，殷坤龙，王佳佳，等. 三峡库区万州区滑坡灾害易发性评价研究 [J]. 岩石力学与工程学报, 2016, 35 (2): 284-296.

[2] 郭靖. 黔西玄武岩地区滑坡易发性评价及玄武岩风化程度判别研究 [D]. 长沙: 中南大学. 2012

[3] 胡畅. 三峡库区秭归—巴东段典型顺层滑坡预测判据研究 [D]. 武汉: 中国地质大学, 2013.

[4] 夏元友. 系统加权聚类法及其在滑坡稳定性预测中的应用 [J]. 自然灾害学报, 1997 (3): 85-91.

[5] 陈黎飞，姜青山，王声瑞. 基于层次划分的最佳聚类数确定方法 [J]. 软件学报, 2008, 19 (1): 62-72

[6] 陆亿红，夏聪. 不确定数据的最优 K 近邻和局部密度聚类算法 [J]. 控制与决策, 2016 (3): 541-546.

[7] 罗清华，彭宇，彭喜元. 一种多维不确定性数据流聚类算法 [J]. 仪器仪表学报, 2013, 34 (6): 1330-1338.

[8] De Carvalho F A T, De Souza R M C R, Chavent M, *et al.* Adaptive Hausdorff distances and dynamic clustering of symbolic interval data [J]. Pattern Recognition Letters, 2006, 27 (3): 167-179.

[9] 杨莉，李南. 软件项目风险应对措施优选的区间模型及其算法 [J]. 控制与决策, 2011, 26 (4): 530-534.

[10] 汪沁，奚李峰. 数据结构 [M]. 北京: 清华大学出版社, 2009

[11] Han Jiawei, Kamber M. 数据挖掘概念与技术 [M]. 范明. 孟小峰, 译. 北京: 机械工业出版社, 2007.

[12] 刘卫明，高晓东，毛伊敏，等. 不确定遗传神经网络在滑坡危险性预测中的研究与应用 [J]. 计算机工程, 2017, 43 (2): 308-316

[13] 毛伊敏，张茂省，程秀娟，等. 基于不确定贝叶斯分类技术的滑坡危险性评价 [J]. 中国矿业大学学报, 2015, 44 (4): 769-774

[14] Yeon Y K, Han J G, Ryu K H. Landslide susceptibility mapping in Injae, Korea, using decision tree. Eng Geol [J]. Engineering Geology, 2010, 116 (3): 274-283.

[15] Guzzetti F, Carrara A, Cardinali M, *et al.* Landslide hazard evaluation: a review of current techniques and their application in a multi-scale study, Central Italy [J]. Geomorphology, 1999, 31 (1-4): 181-216.

[16] 吕启，窦勇，牛新，等. 基于 DBN 模型的遥感图像分类 [J]. 计算机研究与发展, 2014, 51 (9): 1911-1918.

[17] 姜洋，李艳，刘东. 基于 TM 影像属性和形态特征的土地覆被制图方法 [J]. 地球信息科学学报, 2014, 16 (1): 117-125.

[18] 秦鹏程，刘敏，李兰. 有效降水指数在暴雨洪涝监测和评估中的应用 [J]. 中国农业气象, 2016, 37 (1): 84-90.